Q1 What are the 5 V's associated with big data?

Ans1 Volume, Variety, Velocity, Value, Veracity

Q2 There is a file named solution.txt present at the HDFS path 'hdfs:///bigdatads/common\_folder/assignment1/solution.txt', use the command test on the file to perform the following operations:

1. Check whether the path given is in the directory or not
2. Check whether the path given exists or not
3. Check whether the path given is a file or not
4. Check whether the path is empty or not
5. Check whether read permission is granted on the path or not
6. Check whether there is write permission granted on the path or not
7. Check whether the size of the file is 0 or not

Output Format: Paste the commands used to solve this problem in the text box below. Also, attach a screenshot of the code.

Ans 2if $(hdfs dfs -test -d /bigdatads/common\_folder/assignment1/solution.txt) ; then echo "Is a directory";else echo "Is not a directory"; fi

if $(hdfs dfs -test -e /bigdatads/common\_folder/assignment1/solution.txt) ; then echo "Path exists";else echo "Path doesn’t exists"; fi

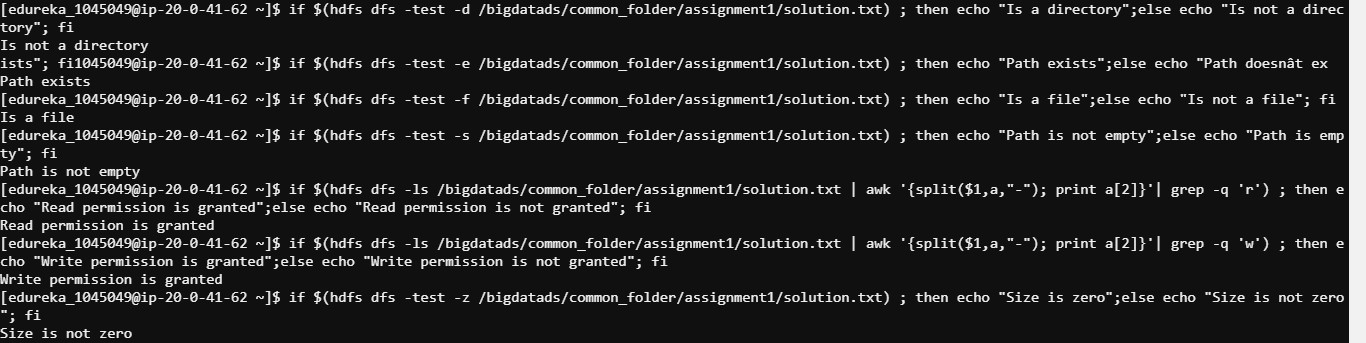
if $(hdfs dfs -test -f /bigdatads/common\_folder/assignment1/solution.txt) ; then echo "Is a file";else echo "Is not a file"; fi

if $(hdfs dfs -test -s /bigdatads/common\_folder/assignment1/solution.txt) ; then echo "Path is not empty";else echo "Path is empty"; fi

if $(hdfs dfs -ls /bigdatads/common\_folder/assignment1/solution.txt | awk '{split($1,a,"-"); print a[2]}'| grep -q 'r') ; then echo "Read permission is granted";else echo "Read permission is not granted"; fi

if $(hdfs dfs -ls /bigdatads/common\_folder/assignment1/solution.txt | awk '{split($1,a,"-"); print a[2]}'| grep -q 'w') ; then echo "Write permission is granted";else echo "Write permission is not granted"; fi

if $(hdfs dfs -test -z /bigdatads/common\_folder/assignment1/solution.txt) ; then echo "Size is zero";else echo "Size is not zero"; fi



Q3 There are three files named patient\_list\_1.csv, patient\_list\_2.csv and patient\_list\_3.csv present at the HDFS path 'hdfs:///bigdatads/common\_folder/assignment1/patient\_list\_x.csv', perform the following with them:

Create a directory covid\_19 in HDFS (Hint: Use Hue)

1. Copy file patient\_list\_1.csv to folder covid\_19
2. Copy file patient\_list\_1.csv to folder covid\_19
3. Copy file patient\_list\_1.csv to folder covid\_19
4. Write a command to merge all these files to a single file named Total\_patients.csv in the folder covid\_19

Output Format: Paste the commands used to solve this problem in the text box below. Also, attach a screenshot of the code.

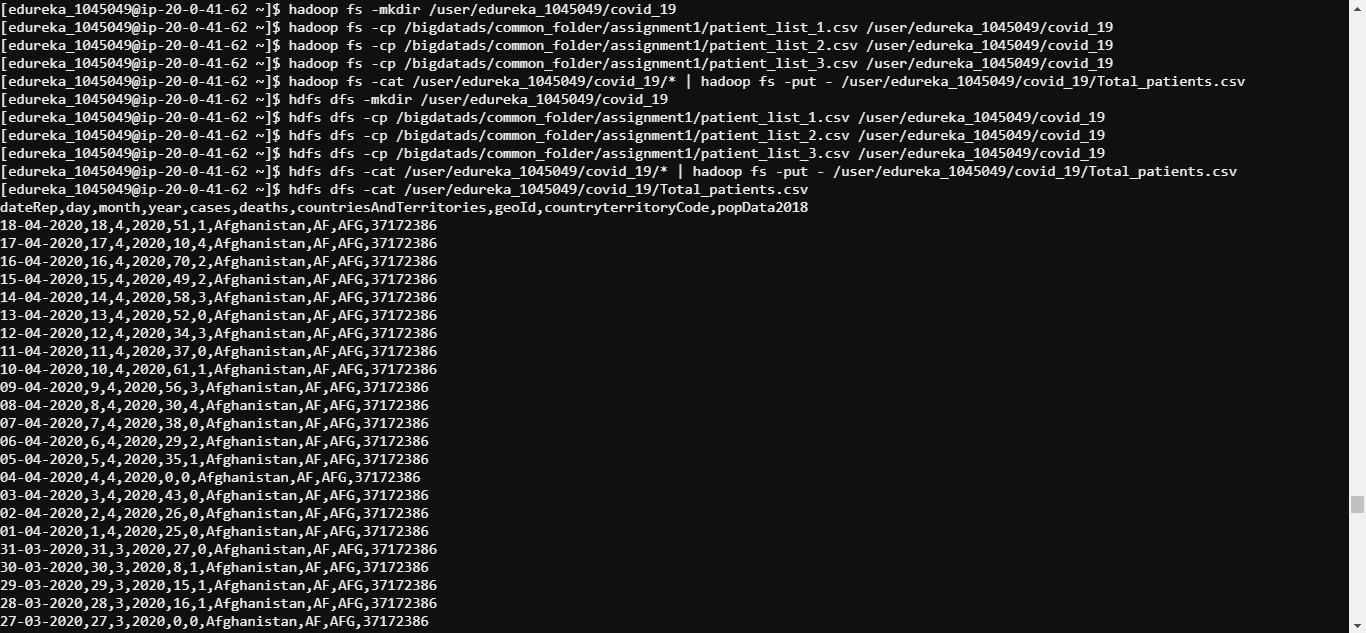
Ans 3 hdfs dfs -mkdir /user/edureka\_1045049/covid\_19

hdfs dfs -cp /bigdatads/common\_folder/assignment1/patient\_list\_1.csv /user/edureka\_1045049/covid\_19

hdfs dfs -cp /bigdatads/common\_folder/assignment1/patient\_list\_2.csv /user/edureka\_1045049/covid\_19

hdfs dfs -cp /bigdatads/common\_folder/assignment1/patient\_list\_3.csv /user/edureka\_1045049/covid\_19

hdfs dfs -cat /user/edureka\_1045049/covid\_19/\* | hadoop fs -put - /user/edureka\_1045049/covid\_19/Total\_patients.csv



Q4 Among the following which one is NOT a storage system/unit?

Ans4 MapReduce

Q5 Which of the following is NOT a streaming data ingestion tool?

Ans5 Apache Storm

Q6 For the file Total\_patients.csv created in question 3, find out the following stats:

1. Size of the file in bytes
2. Group name of the owner of the file
3. Block size for the file
4. Number of replicas of the file present
5. Username of the owner of the file
6. Last modification date of the file

Output Format: Paste the commands used to solve this problem in the text box below. Also, attach a screenshot of the code.

Ans6 hdfs dfs -ls /user/edureka\_1045049/covid\_19/Total\_patients.csv|awk '{print $5}'

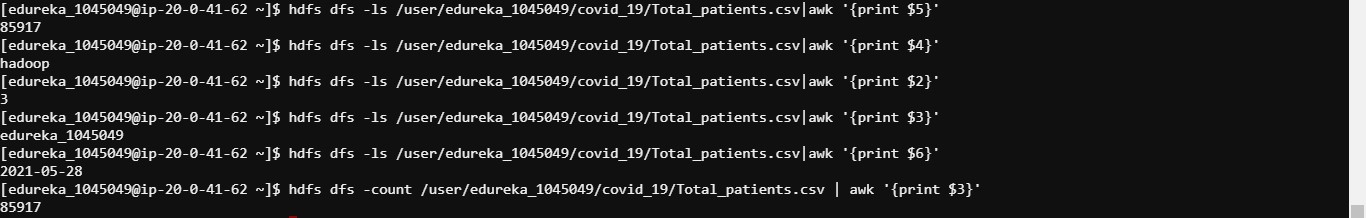
hdfs dfs -ls /user/edureka\_1045049/covid\_19/Total\_patients.csv|awk '{print $4}'

hdfs dfs -ls /user/edureka\_1045049/covid\_19/Total\_patients.csv|awk '{print $2}'

hdfs dfs -ls /user/edureka\_1045049/covid\_19/Total\_patients.csv|awk '{print $3}'

hdfs dfs -ls /user/edureka\_1045049/covid\_19/Total\_patients.csv|awk '{print $6}'

hdfs dfs -count /user/edureka\_1045049/covid\_19/Total\_patients.csv | awk '{print $3}'



Q7. There are three files named file1.txt, file2.txt, file3.txt, and file4.txt present at the HDFS path 'hdfs:///bigdatads/common\_folder/assignment1/filex.txt', perform the following with them:

1. Move these files from your local filesystem to a folder named trash in HDFS
2. Remove all the files from the folder trash

Output Format: Paste the commands used to solve this problem in the text box below. Also, attach a screenshot of the code.

Ans7 hdfs dfs -cp /bigdatads/common\_folder/assignment1/file1.txt /user/edureka\_1045049/

hdfs dfs -cp /bigdatads/common\_folder/assignment1/file2.txt /user/edureka\_1045049/

hdfs dfs -cp /bigdatads/common\_folder/assignment1/file3.txt /user/edureka\_1045049/

hdfs dfs -cp /bigdatads/common\_folder/assignment1/file4.txt /user/edureka\_1045049/

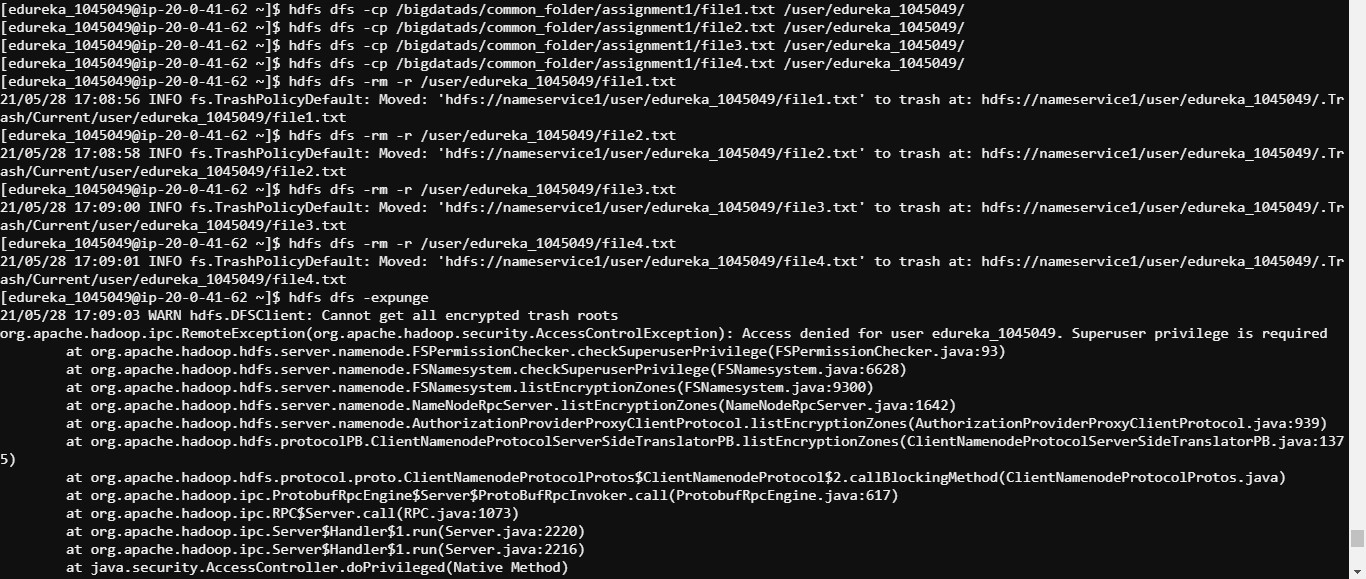
hdfs dfs -rm -r /user/edureka\_1045049/file1.txt

hdfs dfs -rm -r /user/edureka\_1045049/file2.txt

hdfs dfs -rm -r /user/edureka\_1045049/file3.txt

hdfs dfs -rm -r /user/edureka\_1045049/file4.txt

hdfs dfs -expunge



Q8. From the following components of Yarn which one gets initiated per application basis?

Ans8 Resource manager

Q9. When the resource manager approves the request for resources requested by application master, the job gets launched by application master in \_\_\_\_\_\_.

Ans9. Container

Q10. There is a file named geological\_data.csv present at the HDFS path 'hdfs:///bigdatads/common\_folder/assignment1/geological\_data.csv', perform the following on the file:

1. Copy the file geological\_data.csv from local to the folder geological in HDFS
2. After the file has been copied, view the contents of the file
3. If the file has “old stuff” written at its bottom then remove the file from HDFS, if not then keep it as it is and change the replication factor of the file to 1

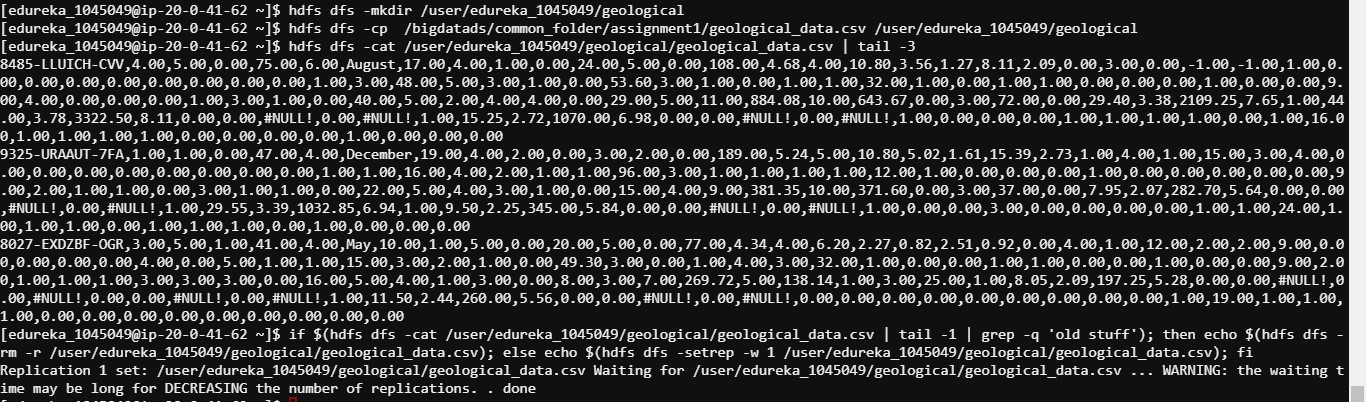
Output Format: Paste the commands used to solve this problem in the text box below. Also, attach a screenshot of the code.

Ans10. hdfs dfs -mkdir /user/edureka\_1045049/geological

hdfs dfs -cp /bigdatads/common\_folder/assignment1/geological\_data.csv /user/edureka\_1045049/geological

hdfs dfs -cat /user/edureka\_1045049/geological/geological\_data.csv | tail -3

if $(hdfs dfs -cat /user/edureka\_1045049/geological/geological\_data.csv | tail -1 | grep -q 'old stuff'); then echo $(hdfs dfs -rm -r /user/edureka\_1045049/geological/geological\_data.csv); else echo $(hdfs dfs -setrep -w 1 /user/edureka\_1045049/geological/geological\_data.csv); fi



Q11 How is Apache Hadoop better than RDBMS?

Ans11 Works better on unstructured and semi-structured data

Q12 Which of the files contains the configuration setting for HDFS daemons?

Ans12 hdfs-site.xml

Q13 Write the following commands:

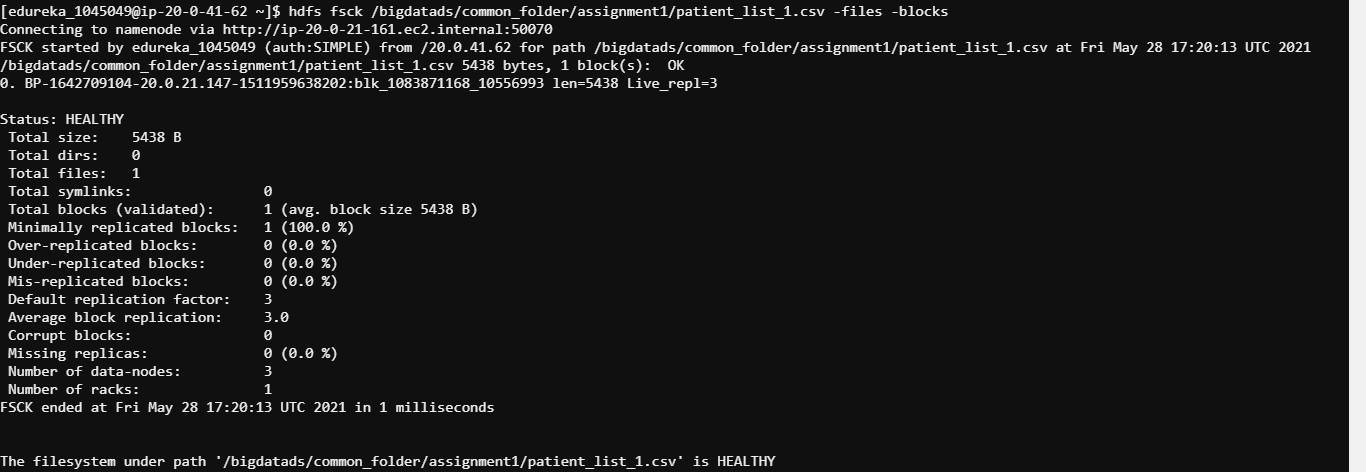
1. Check the capacity, size, and available free space on the HDFS file system and display the content in a human-readable format
2. Check the health of the files patient\_list\_1.csv, patient\_list\_2.csv and patient\_list\_3.csv, present in the directory assignments, used in question 3

Output Format: Paste the commands used to solve this problem in the text box below. Also, attach a screenshot of the code.

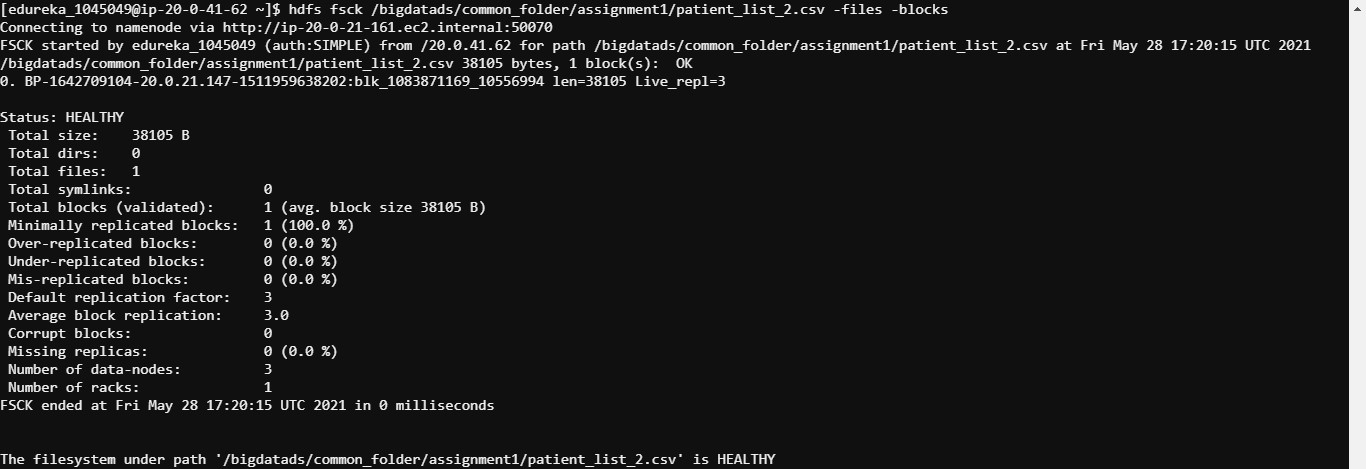
Ans13 hdfs dfs -df -h



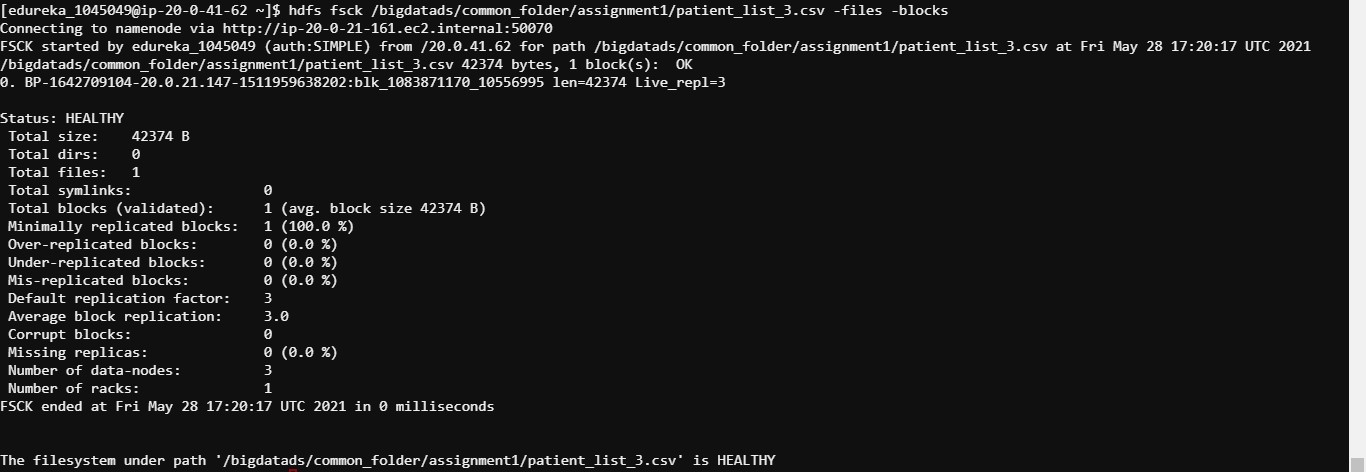
hdfs fsck /bigdatads/common\_folder/assignment1/patient\_list\_1.csv -files -blocks



hdfs fsck /bigdatads/common\_folder/assignment1/patient\_list\_2.csv -files -blocks



hdfs fsck /bigdatads/common\_folder/assignment1/patient\_list\_3.csv -files -blocks



Q14 The property of switching from a passive name node to an active name node in the Hadoop cluster is known as:

Ans14 High Availability